

Requested Patent: FR2803928A1

Title:

PROCESSING OF NATURAL LANGUAGE TEXT TO EVALUATE THE CONTENT  
FOR MARKING IN AN EDUCATIONAL CONTEXT, USES COMPARISON OF  
ENTERED TEXT TO SET OF STORED KEY WORDS TO DETERMINE SCORE

Abstracted Patent: FR2803928

Publication Date: 2001-07-20

Inventor(s): MULLER BERNARD GASTON FRANCOIS

Applicant(s): AURALOG (FR)

Application Number: FR20000000590 20000118

Priority Number(s): FR20000000590 20000118

IPC Classification: G06F17/28

Equivalents:

**ABSTRACT:**

The processing system has an interface to allow the user to introduce their response in text form. The processing makes use of a stored (13) list of key words associated with the question posed, and compares (11) the text with key words to detect coincidence. A score (14) is computed based on the coincidence, and made available for further processing.

(19) RÉPUBLIQUE FRANÇAISE  
INSTITUT NATIONAL  
DE LA PROPRIÉTÉ INDUSTRIELLE  
PARIS

(11) N° de publication :  
(à n'utiliser que pour les  
commandes de reproduction)

2 803 928

(21) N° d'enregistrement national : 00 00590

(51) Int Cl<sup>7</sup> : G 06 F 17/28

(12)

## DEMANDE DE BREVET D'INVENTION

A1

(22) Date de dépôt : 18.01.00.

(30) Priorité :

(43) Date de mise à la disposition du public de la demande : 20.07.01 Bulletin 01/29.

(56) Liste des documents cités dans le rapport de recherche préliminaire : Se reporter à la fin du présent fascicule

(60) Références à d'autres documents nationaux apparentés :

(71) Demandeur(s) : AURALOG Société anonyme — FR.

(72) Inventeur(s) : MULLER BERNARD GASTON FRANCOIS.

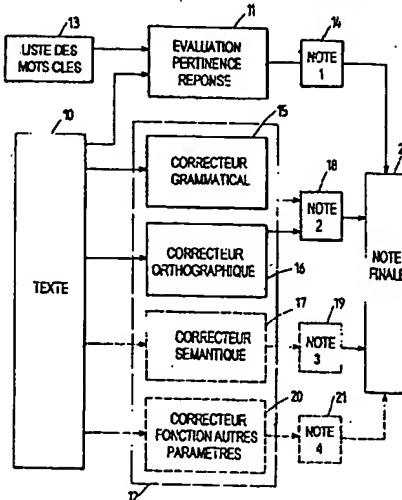
(73) Titulaire(s) :

(74) Mandataire(s) : CABINET DE BOISSE ET COLAS.

(54) SYSTEME DE TRAITEMENT DE DONNEES POUR L'EVALUATION D'UN TEXTE.

(57) Ce système de traitement de données pour l'évaluation d'un texte en langage naturel élaboré par un utilisateur en réponse à une consigne transmise audit utilisateur comprend :

- des premiers moyens d'interface pour l'introduction par ledit utilisateur dudit texte en réponse à ladite consigne, et
- des moyens de traitement de données (11) pour l'évaluation dudit texte (10), comprenant:
  - \* des moyens (13) de mémorisation d'au moins une liste de mots-clés associés à ladite consigne,
  - \* des moyens de comparaison (11) pour identifier des mots dudit texte (10) coïncidant avec des mots de ladite liste de mots-clés mémorisés, et
  - \* des moyens de calcul (11) pour générer une donnée (14) d'évaluation dudit texte en fonction du résultat de ladite comparaison.



FR 2 803 928 - A1



L'invention a pour objet un système de traitement de données pour l'évaluation d'un texte.

Dans le domaine de l'enseignement, notamment l'apprentissage des langues, l'évaluation de la qualité et de la pertinence d'un texte écrit en langage naturel en réponse à une consigne ou une question repose sur une appréciation portée par une personne physique, le correcteur.

Par ailleurs, il existe des logiciels éducatifs, notamment d'apprentissage des langues, qui, au moyen d'un ordinateur personnel, permettent à un individu d'effectuer un apprentissage sans intervention d'un enseignant ou correcteur personne physique. Ces systèmes permettent de corriger automatiquement les réponses faites par l'utilisateur à des questions qui lui sont posées dans le cas où il existe un nombre fini et limité de réponses possibles à ces questions.

L'invention vise à fournir un système de traitement de données permettant d'évaluer automatiquement la pertinence d'un texte rédigé librement en langage naturel en réponse à une consigne. On entend ici par consigne l'ensemble des indications données à un utilisateur du système pour rédiger son texte. Cette consigne peut se présenter par exemple sous la forme d'un texte (tel qu'une question, un thème, etc.) associé à un document (tel qu'un enregistrement audio, une image fixe, une image vidéo, etc.). Cette consigne est présentée à l'utilisateur par des moyens d'interface tels que transducteur acoustique, écran d'affichage, écran vidéo, etc.

A cet effet, l'invention a pour objet un système de traitement de données pour l'évaluation d'un texte en langage naturel élaboré par un utilisateur en réponse à une consigne transmise audit utilisateur, comprenant :

- des premiers moyens d'interface pour l'introduction par ledit utilisateur dudit texte en réponse à ladite consigne, et

5 - des moyens de traitement de données pour l'évaluation dudit texte,

caractérisé en ce que lesdits moyens de traitement de données comprennent :

10 \* des moyens de mémorisation d'au moins une liste de mots-clés associés à ladite consigne,

\* des moyens de comparaison pour identifier des mots dudit texte coïncidant avec des mots de ladite liste de mots-clés mémorisés, et

\* des moyens de calcul pour générer une donnée d'évaluation dudit texte en fonction du 15 résultat de ladite comparaison.

De préférence, le système selon l'invention comporte en outre une ou plusieurs des caractéristiques suivantes considérées seules ou en combinaison :

20 - lesdits moyens de mémorisation contiennent plusieurs listes de mots-clés, chaque liste regroupant un ensemble de mots-clés affectés à un concept associé à ladite consigne ;

25 - lesdits moyens de calcul sont adaptés pour calculer une note fonction du nombre de listes dont ledit texte contient au moins un mot-clé ;

30 - lesdits moyens de mémorisation contiennent un coefficient de pertinence affecté à chacun desdits mots-clés et lesdits moyens de calcul sont adaptés pour calculer ladite note en fonction du coefficient de pertinence d'au moins une partie des mots-clés contenus dans le texte ;

- à chaque liste est affectée une valeur maximale et lesdits moyens de calcul sont adaptés pour :

35 \* calculer pour chaque liste une valeur pondérée fonction du coefficient de pertinence d'au

moins un mot-clé de ladite liste contenu dans ledit texte,

\* calculer ladite note en fonction de la somme des valeurs pondérées desdites listes ;

5 - le système comprend des moyens de vérification orthographique et/ou grammaticale et/ou sémantique dudit texte et lesdits moyens de calcul sont adaptés pour générer ladite donnée d'évaluation en fonction des résultats de ladite vérification et de ladite comparaison ;

10 - lesdits moyens de calcul sont adaptés pour calculer une note de qualité fonction du nombre de fautes détectées dans ledit texte par lesdits moyens de vérification et une note de pertinence fonction du résultat de ladite comparaison ;

15 - lesdits moyens de calcul sont adaptés pour générer ladite donnée d'évaluation en fonction desdites notes de qualité et de pertinence ;

- le système comporte des moyens de génération de ladite consigne et, pour la transmission de ladite consigne audit utilisateur, des seconds moyens d'interface comprenant au moins l'un parmi les moyens de :

\* affichage alphanumérique,

\* affichage graphique,

\* affichage vidéo,

25 \* reproduction de messages audio ;

- lesdits moyens d'introduction dudit texte comprennent au moins l'un parmi plusieurs moyens comprenant un clavier, des moyens de reconnaissance de l'écriture, des moyens de reconnaissance vocale.

30 L'invention vise également l'application d'un système de traitement de données tel que défini ci-dessus à l'apprentissage des langues étrangères.

D'autres caractéristiques et avantages de l'invention ressortiront de la description qui va suivre,

35 faite en se référant aux dessin annexés sur lesquels :

- la figure 1 est un schéma-bloc matériel simplifié d'un exemple de réalisation du système de traitement de données selon l'invention basé sur un ordinateur personnel, et

5 - la figure 2 est un schéma-bloc fonctionnel illustrant les fonctions mises en oeuvre dans le système de traitement de données selon l'invention.

10 Selon l'exemple de réalisation de la figure 1, le système de traitement de données selon l'invention est basé sur un ordinateur personnel (PC) 1 convenablement programmé. De manière essentielle, le PC 1 est équipé de moyens de traitement de données 2 (microprocesseur) et de mémoires 3, ainsi que d'un certain nombre d'interfaces. Ces interfaces comprennent un écran d'affichage 4, un 15 clavier 5, un dispositif 6 d'acquisition des données et programmes nécessaires à l'exécution des fonctions qui sont décrites dans la suite, et facultativement un ou plusieurs transducteurs électro-acoustiques (HP) 7. Le dispositif 6 peut être constitué, par exemple, par un 20 lecteur de disquette, CD-ROM, DVD-ROM ou autre moyen de stockage de données. Il peut s'agir également d'un dispositif d'échange de données au moyen duquel l'ordinateur personnel 1 se trouve relié par un réseau de communications tel qu'un réseau local ou Internet à un 25 serveur à partir duquel les programmes précités, ou une partie de ceux-ci, sont téléchargés.

Il s'agit là d'un simple exemple de réalisation du système de traitement de données suivant l'invention et celui-ci pourrait revêtir d'autres formes, par exemple 30 celle d'un ordinateur central contenant les programmes précités et auxquels l'utilisateur a accès via un terminal.

On se reporterà également à la figure 2 sur laquelle sont explicitées les fonctions mises en oeuvre par le 35 système de traitement de données de la figure 1.

Lorsqu'un utilisateur a accédé sur son PC 1 à l'application concernée, stockée par exemple sur un CD ROM 8 lu par le lecteur 6, il lui est présentée une consigne l'invitant à rédiger un texte en langage naturel.

Cette consigne consiste en des indications relatives au sujet ou thème du texte que doit élaborer l'utilisateur. Cette consigne peut se présenter sous la forme d'une ou plusieurs questions, d'un texte définissant le sujet ou thème à traiter, d'une image fixe, d'une séquence vidéo ou d'une combinaison d'un ou plusieurs de ces média. Cette consigne est présentée à l'utilisateur via les moyens d'interface du PC tels que l'écran 4 et/ou le transducteur électro-acoustique 7.

En réponse à la consigne, l'utilisateur élabore un texte en langage naturel et l'introduit dans le PC 1 au moyen du clavier 5. En variante, le texte pourrait être introduit dans le PC 1 par d'autres moyens d'interface non représentés à la figure 1, par exemple oralement via un microphone et des moyens de reconnaissance vocale, ou sous forme manuscrite via une ardoise électronique et des moyens de reconnaissance d'écriture.

Le texte 10 introduit dans le PC 1 est soumis respectivement en 11 à un processus d'évaluation de sa pertinence et en 12 à un processus d'évaluation de sa qualité.

Le processus 11 d'évaluation de la pertinence du texte repose sur le stockage d'un certain nombre de mots-clés qui sont associés à la consigne et qui permettent de vérifier l'adéquation de la réponse (le texte 10) à la question et/ou le document de référence (la consigne). De préférence, ces mots-clés sont organisés en un certain nombre de listes correspondant respectivement à différents concepts-clés associés à la consigne. Chaque concept-clé est ainsi défini par une liste de mots qui

illustrent le concept. Le terme liste doit être compris dans une acception large comme désignant un ensemble de mots-clés stockés en mémoire avec un lien les rattachant entre eux et les distinguant des mots-clés d'autres 5 ensembles ou liste.

En outre, un coefficient de pertinence est de préférence associé à chaque mot-clé, pour rendre compte de sa proximité sémantique avec le concept-clé correspondant. Par exemple, au mot "maison" pourraient 10 être associés les mots "maison", "chalet", "appartement" avec un coefficient de pertinence de valeur maximale (valeur 1 par exemple), et les mots "hutte", "suite", "habitat", "habitation", "caserne", "château", ... avec 15 un coefficient de pertinence de valeur plus faible (compris entre 0 et 1 par exemple). Les mots définissant un concept-clé dépendent évidemment du concept, mais ils peuvent être également liés au contexte d'utilisation de celui-ci selon la question posée, le document de référence utilisé... Les listes de mots-clés et les 20 coefficients de pertinence qui leur sont associés sont élaborés par les concepteurs de l'application et stockés en mémoire comme indiqué précédemment et comme illustré par la référence 13 à la figure 2.

Ainsi, au niveau du bloc 11, les moyens de 25 traitement de données 2 procèdent à une comparaison entre les mots du texte 10 et ceux des listes 12 de mots-clés. A partir de cette comparaison, les moyens de traitement 2 calculent une note de pertinence qui est fonction du nombre de listes ou concepts-clés dont au moins un mot-clé est contenu dans le texte 10. Le mode de calcul de la 30 note peut être adapté en fonction des besoins.

A titre d'exemple, si le texte 10 contient plusieurs mots-clés d'une même liste, il peut être choisi de ne retenir que celui ayant le coefficient de pertinence de 35 valeur la plus élevée. La valeur maximale de chaque liste

ou concept-clé peut être fixée à 1 et le coefficient de pertinence des mots-clés compris entre 0 et 1. La note attribuée à la pertinence du texte 10 sera donc alors constituée de la somme des coefficients de pertinence des différents mots-clés retenus dans chaque liste (à savoir, dans cet exemple, un seul par liste) rapportés au nombre de listes ou concepts-clés.

En outre, il peut être prévu dans les listes de mots-clés des mots pénalisants, c'est-à-dire des mots qui ne devraient pas être utilisés dans le texte 10 compte tenu de la consigne et de son contexte, par exemple de faux amis. De préférence, ces mots sont affectés de coefficients de pondération négatifs et viennent donc, lorsqu'ils se rencontrent dans le texte 10, diminuer la note élaborée en 11. Celle-ci est désignée Note 1 au bloc 14.

Parallèlement au processus d'évaluation de la pertinence du texte en 11, il est procédé en 12 à l'évaluation de sa qualité au moyen d'un correcteur grammatical 15, d'un correcteur orthographique 16 et, facultativement, d'un correcteur sémantique 17.

Un correcteur orthographique est un logiciel permettant d'indiquer, dans un texte quelconque, tous les mots qui ne figurent pas dans un dictionnaire de référence. Idéalement, ce dictionnaire contient tous les mots, avec leurs déclinaisons, existant dans la langue du texte.

Un correcteur grammatical est un logiciel permettant d'indiquer si un texte est grammaticalement correct et, le cas échéant, d'indiquer où se situent les erreurs et la nature de celles-ci. Les erreurs peuvent concerner par exemple les accords, la formation des phrases, le respect des règles de grammaire, etc.

En pratique, un correcteur grammatical intègre un correcteur orthographique, mais ils ont été représentés

sous forme séparée sur le dessin pour des raisons de clarté.

Les correcteurs orthographiques et grammaticaux sont largement utilisés en association avec les logiciels de traitement de texte les plus connus et ne seront donc pas décrits. On citera pour mémoire les produits suivants :

- CORRECT ENGLISH, de la société Lernout et Hauspie (Belgique) pour la langue anglaise,

10 - CORRECTEUR 101 et EL CORRECTOR de la société Machina Sapiens (Canada) pour les langues française et espagnole respectivement ;

- ERRATA CORRIGE de la société Expert Systems (Italie) pour la langue italienne.

15 Le correcteur grammatical 15 et le correcteur orthographique 16 permettent de détecter des fautes dans le texte 10, et de calculer en 18 une seconde note, désignée Note 2, en fonction du nombre de fautes détectées.

20 Dans le système décrit dans la présente demande, le correcteur grammatical 15 et le correcteur orthographique 16 sont utilisés essentiellement à des fins de vérification pour noter la qualité du texte 10. Bien entendu, ces logiciels peuvent, au choix du concepteur de l'application, être utilisés également dans leur fonction de correcteurs en présentant à l'auteur du texte les fautes orthographiques et grammaticales qu'il a commises, par exemple par voie d'affichage dans le texte considéré.

25 De manière facultative, une troisième note, désignée Note 3 au bloc 19, peut être élaborée au moyen du correcteur sémantique 17. Un correcteur sémantique est un logiciel permettant de vérifier la cohérence sémantique du texte analysé. Il permet par exemple de rejeter les phrases grammaticalement correctes, mais absurdes, telles que par exemple "la carotte dévore le lapin". En 30 variante, le nombre de fautes détectées par le correcteur

sémantique 17 peut constituer un paramètre de calcul de la note 2 en 18 au lieu de donner lieu au calcul d'une note séparée en 19 comme représenté à la figure 2.

D'autres paramètres tels que le nombre de mots du 5 texte 10, la longueur moyenne des phrases, le temps mis par l'utilisateur pour formuler sa réponse (texte 10), peuvent également être pris en compte en 20 pour calculer une quatrième note désignée Note 4 en 21.

Enfin, à partir des notes calculées en 14, 18 et 10 éventuellement 19 et 21, les moyens de traitement 2 calculent en 22 une Note finale qui est la note globale d'évaluation de la qualité et de la pertinence du texte 10. Cette Note finale est communiquée à l'auteur du texte 10, par exemple par affichage sur l'écran 4. 15 Naturellement, il est également possible d'afficher pour l'auteur du texte 10 les notes individuelles calculées en 14, 18 et éventuellement 19 et 21.

La Note finale du bloc 22 peut se présenter soit sous la forme d'un nombre de points rapportés à une 20 valeur maximale, soit comme un degré dans une échelle de notation, soit encore sous n'importe quelle forme appropriée. Le calcul de la Note finale en 22 peut bien entendu faire appel à des coefficients appliqués aux notes des blocs 14, 18, 19 et 21. De même, de tels 25 coefficients peuvent être appliqués par le correcteur grammatical 15, le correcteur orthographique 16 et le correcteur sémantique 17 en fonction de la gravité des fautes détectées.

A titre d'exemple, si on veut donner une note 30 globale sur 20 (en supposant qu'il n'y a pas de correcteur sémantique 17 et de correcteur 20 en fonction d'autres paramètres), on peut noter sur 10 le résultat du correcteur grammatical et orthographique 15, 16 (10 - le nombre de fautes détectées) et sur la 10 la présence des 35 concepts-clés (si cinq concepts-clés ou listes de mots-

clés sont définis pour une consigne, on peut attribuer deux points pour chaque concept-clé retrouvé dans le texte, le coefficient de pertinence des mots utilisés servant à moduler l'attribution de ces points).

5 Le système de traitement décrit peut être utilisé par exemple pour la mise en oeuvre d'un logiciel multimédia d'apprentissage des langues étrangères. L'utilisateur prend connaissance d'un document (par exemple une photo ou un texte affiché sur l'écran 4) et 10 répond à une question ou indication relative à ce document (par exemple : "décrivez la photo", "résumer le texte", etc.). La consigne peut comprendre des directives, par exemple quant au nombre maximal de mots que doit contenir le texte. L'utilisateur introduit 15 celui-ci dans le PC 1, par exemple au moyen du clavier 5, et lorsqu'il a validé définitivement ce texte, il se voit attribuer une note finale comme décrit ci-dessus en regard de la figure 2.

Il va de soi que le mode de réalisation décrit n'est 20 qu'un exemple et l'on pourrait le modifier, notamment par substitution d'équivalents techniques, sans sortir pour cela du cadre de l'invention.

## REVENDICATIONS

1. Système de traitement de données pour l'évaluation d'un texte en langage naturel élaboré par un utilisateur en réponse à une consigne transmise audit 5 utilisateur, comprenant :

- des premiers moyens d'interface pour l'introduction par ledit utilisateur dudit texte en réponse à ladite consigne, et

10 - des moyens de traitement de données pour l'évaluation dudit texte,

caractérisé en ce que lesdits moyens de traitement de données (2) comprennent :

15 \* des moyens (8) de mémorisation d'au moins une liste (13) de mots-clés associés à ladite consigne,

\* des moyens de comparaison (2) pour identifier des mots dudit texte (10) coïncidant avec des mots de ladite liste de mots-clés mémorisés, et

20 \* des moyens de calcul (2) pour générer une donnée (14,22) d'évaluation dudit texte en fonction du résultat de ladite comparaison.

25 2. Système selon la revendication 1, caractérisé en ce que lesdits moyens de mémorisation (8) contiennent plusieurs listes de mots-clés (13), chaque liste regroupant un ensemble de mots-clés affectés à un concept associé à ladite consigne.

30 3. Système selon la revendication 2, caractérisé en ce que lesdits moyens de calcul (2) sont adaptés pour calculer une note (14) fonction du nombre de listes dont ledit texte contient au moins un mot-clé.

35 4. Système selon la revendication 3, caractérisé en ce que lesdits moyens de mémorisation (8) contiennent un coefficient de pertinence affecté à chacun desdits mots-clés et en ce que lesdits moyens de calcul (2) sont adaptés pour calculer ladite note (14) en fonction du

coefficient de pertinence d'au moins une partie des mots-clés contenus dans ledit texte (10).

5. Système selon la revendication 4, caractérisé en ce qu'à chaque liste est affectée une valeur maximale et 5 en ce que lesdits moyens de calcul (2) sont adaptés pour :

- \* calculer pour chaque liste une valeur pondérée fonction du coefficient de pertinence d'au moins un mot-clé de ladite liste contenu dans ledit texte,
- 10 \* calculer ladite note (14) en fonction de la somme des valeurs pondérées desdites listes.

6. Système selon l'une quelconque des revendications 1 à 5, caractérisé en ce qu'il comprend des moyens (15,16,17) de vérification orthographique et/ou 15 grammaticale et/ou sémantique dudit texte et en ce que lesdits moyens de calcul (2) sont adaptés pour générer ladite donnée d'évaluation (22) en fonction de ladite vérification et de ladite comparaison.

20 7. Système selon la revendication 6, caractérisé en ce que lesdits moyens de calcul (2) sont adaptés pour calculer une note de qualité (18,19) fonction du nombre de fautes détectées dans ledit texte par lesdits moyens de vérification et une note de pertinence (14) fonction du résultat de ladite comparaison.

25 8. Système selon la revendication 7, caractérisé en ce que lesdits moyens de calcul (2) sont adaptés pour générer ladite donnée d'évaluation (22) en fonction desdites notes de qualité (18,19) et de pertinence (14).

30 9. Système selon l'une quelconque des revendications 1 à 8, caractérisé en ce qu'il comporte des moyens (8) de génération de ladite consigne et, pour la transmission de ladite consigne audit utilisateur, des seconds moyens d'interface (4,7) comprenant au moins l'un parmi les moyens de :

- 35 \* affichage alphanumérique,

- \* affichage graphique,
- \* affichage vidéo,
- \* reproduction de messages audio.

10. Système selon l'une quelconque des  
5 revendications 1 à 9, caractérisé en ce que lesdits moyens (5) d'introduction dudit texte comprennent au moins l'un parmi plusieurs moyens comprenant un clavier, des moyens de reconnaissance de l'écriture, des moyens de reconnaissance vocale.

10 11. Système selon l'une quelconque des revendications 1 à 10, caractérisé en ce qu'il est appliqué à l'apprentissage des langues étrangères.

1/1

FIG.:1

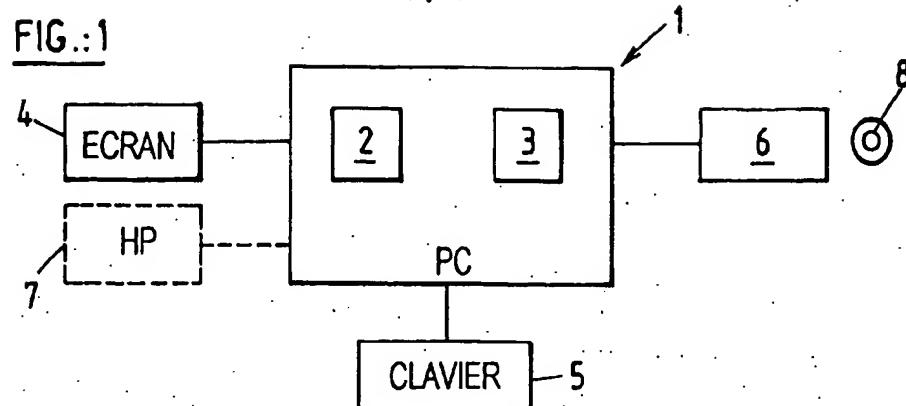
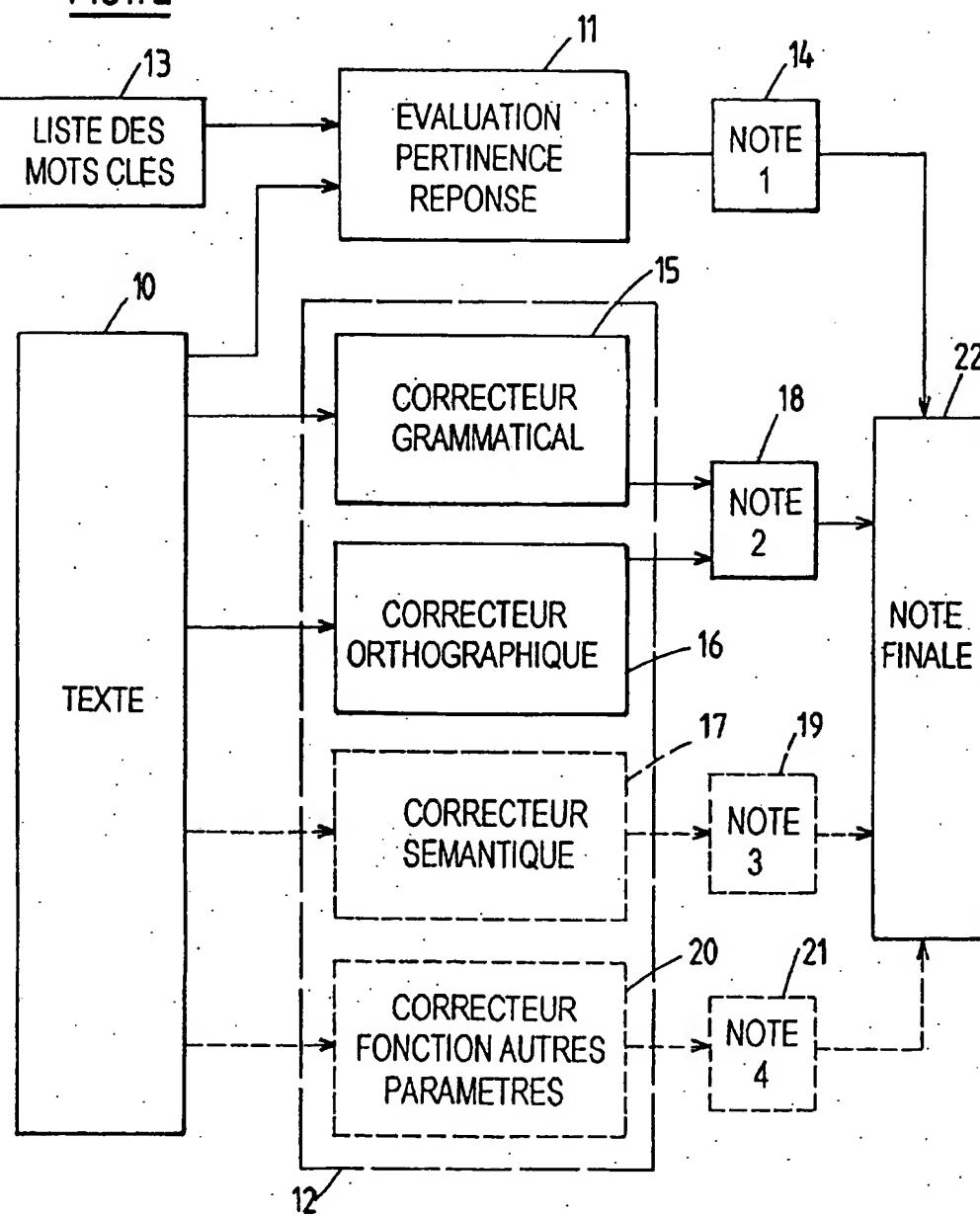


FIG.:2



RAPPORT DE RECHERCHE  
PRÉLIMINAIREétabli sur la base des dernières revendications  
déposées avant le commencement de la recherche

2803928

N° d'enregistrement  
nationalFA 587810  
FR 0000590

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	US 4 730 270 A (OKAJIMA ATSUSHI ET AL) 8 mars 1988 (1988-03-08) Document en entier	1-11	G06F17/28
X	US 5 727 950 A (PADWA DAVID J ET AL) 17 mars 1998 (1998-03-17) * colonne 42, ligne 63 - colonne 64, ligne 25 * Document en entier	1-10	
X	WO 97 08604 A (LIDDY ELIZABETH D ; LI MING (US); PAIK WOOGIN (US); UNIV SYRACUSE ()) 6 mars 1997 (1997-03-06) * page 1-3 *	1-10	
A	US 5 820 386 A (SHEPPARD II CHARLES BRADFORD) 13 octobre 1998 (1998-10-13) * colonne 5, ligne 20 - colonne 7, ligne 66 *	1-11	
		DOMAINES TECHNIQUES RECHERCHÉS (Int.CL7)	
		G09B	
1		Date d'achèvement de la recherche	
15 septembre 2000		Examinateur	
CATÉGORIE DES DOCUMENTS CITÉS			
X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : antérieur-plan technologique O : divulgation non-écrite P : document intercalaire		T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant	